On the design of a statistical database, micro-, macro- and metadata modelling

Lenz, Hans-J.

Historical Social Research, Vol. 18 — 1993 — No. 4, 31-48

On the Design of a Statistical Database, Micro-, Macro- and Metadata Modelling

*Hans - J. Lenz**

Abstract: Statistical databases have some non-standard characteristics which cannot be well supported by commercially available database management systems. This concerns the underlying data structures, the various abstraction levels of the data, the type of operations on the data and the kind of processing requirements. The current research is aiming at an appropriate conceptual modelling, an efficient representation of the data and a powerful and user-friendly query processing on each data level, i.e. micro-, macro- and metadata. It's main results are published in the proceedings of the biannual conferences on »Statistical and Scientific Database Management, in some issues of the statistical and database journals, in recently published textbooks and in proceedings published by Eurostat since 1992 as part of the DOSES program. In the following we describe the state of the art of the statistical database design and cover some future trends in this research area. The reasoning is mainly influenced by ideas of the »American« *(Shoshani),* »Italian« *(Rafanelli, Ricci),* the »Japanese« *(Sato)* and the »Swedish« *(Sundgren)* school. Moreover, the continuing experience from the collaboration with the Statistical Office of Berlin on the new database system of the city of Berlin can be considered to be a prerequisite for reasearch in this field.

1. Introduction

A database is called a statistical database (StDB) if it contains data of three kinds:

- **Microdata** as primary or basis data on individuals, objects or events representing sampled, census or collected data.

* Address all communications to Hans-J. Lenz, Freie Universität Berlin, Institut für Wirtschaftsinformatik und Operations Research, Garystr. 21, D-14195 Berlin.

- **Macrodata** as grouped or aggregated data (summarized data) which are cross-classified by a set of categorical attributes(variables). The summary attribute represents counts(frequencies), means, indices orother statistics characterizing a set(population) of individuals, objects or events.
- **Metadata** describing the micro- and macrodata on the semantic, structural, statistical and physical level in such a way that they can be stored, transformed retrieved and transmitted in a reasonable way. It covers the whole »data life cycle«, i.e. the data collecting from the data source, the data storing, the data processing and retrieval, and the data disseminating within the electronic data interchange (EDI). As a matter of fact, the metadata itself must be easily accessible similiar to the micro- and macrodata.

We close the introduction with some examples of typical retrieval operations (queries) on the data of the above kind. We make use of a pseudo-code notation. In some cases the system response (»i^«) is given.

- Microdata
  list name, age, sex
  from labourcensusemployees
  where industry = 'whole industry' **and** year = 1980

- **Macrodata**
  list number (employees), **average** (employees.income)
  from labourcensus
  where industry - 'whole industry' **and** year - 1980
  cross-classified by age^group **and** sex

- Metadata
  ? household
  ☞ *All the people belong to a household who live there together and have a joint budget Each person who has an own budget forms her own household,*
  ? summary-attribute (employees)
  ☞ income
  ? categoryattribute (employees)
  ? domain (industry)

2. Special features of a statistical database

The data structures offered by conventional database management systems are rather inconvenient for storing and retrieving statistical data. Moreover, the functionality of the systems is inappropriate and inadequate for interpreting, validating and analyzing such kind of data. The main reasons are the following, cf. Shoshani (1991):

- *Set- (tupel- or row-)oriented as well as ordered structures must berepresented.*
- *Vectors, matrices, nested arrays besides non-fixed formatted data types like documents, images, plots must be stored.*
- *Almost static (historical) micro- and macrodata with very slow update rates exist.*
- *GBytes of data are to be stored some of which are sparse.*
- *Very complex semantic integrity constraints are defined on the micro- and macrodata, espescially in order to map their links.*
- *Micro- and macrodata without the corresponding metadata are useless.*
- *The locality (^clusterings) of data according to attributes (columns) or records (rows) is context-dependent.*
- *The querying and the processing of statistical data can cause »long transactions«.*
- *The creation of macrodata by grouping, aggregation etc. implies the bookkeeping of metadata.*
- *The frontier (interface) between retrieval and statistical analysis is not crisp.*
- *The retrieval of even »anonymous« data must care about privacy rules.*

The above special features of a statistical database (STDB) give raise to the following requirements for the design of a statistical database:

- *The conceptual data model must include the micro-, macro- and metadata level.*
- *The data-structures for micro-, macro- and metadata must be efficient.*
- *The set of operators should be complete which is defined on the micro-, macro- and meta data level.*
- *The user-interfaces for the different user-groups must be user-friendly.*
- *Privacy handling mechanism must be incorporated.*

3. Some Statistical Data Models

Following *Ullman (1988)* a data model is a notation for describing data and a set of operations used to manipulate the data. Many proposals for such models have been published where most of them use a graph-theoretic approach. The most prominent models are presented in the next sub-chapters. A common feature of these models is the fact that most of them deal only with macro- and metadata. The macrodata are linked to a specific topic (context) or a field of interest Several statistical populations are part of such a field of interest. A specific population is considered as a set or class of units (instances) each of which are characterized by a set of properties (attributes) which are related to each other. The attributes can be classified *(Shoshani, 1982)* according to their

role. Attributes used to categorize or to index data are called category attributes and the attributes expressing summary properties are called summary attributes. The value of a category attribute is sometimes called »category« instead of »category value«.

3.1 *SUBJECT*

Chan, Shoshani (1981) consider a single multi-dimensional table which is modelled as a root tree (V, E) where the set V of nodes includes

C-nodes (cluster-nodes): a cluster is a set of subordinate categories, e.g. the node »Year« has subordinates 1980, 1985, 1990.

X-nodes (cross-product nodes): a cross product is the Cartesian product of sets of categories,e.g. domain (Sex) **x** domain (Year); note that the root node is of this type and refers to a complete table.

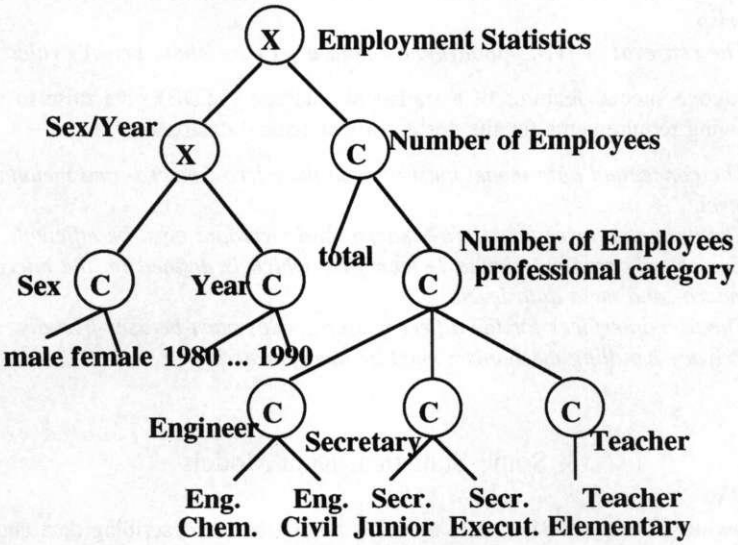

Fig. 1: A ***Subject*** tree

The left subtree represents the stub of a statistical table and is modelled by C- and X-nodes.The right subtree models the heading of the table and consists of C- and X-nodes. The different summary attributes are clustered in a C-node,

indicating that the subordinate nodes are pointing to data columns in the multi-dimensional table. Besides of being marked with either a »C« or a »X« , the nodes are labeled according to the role they play in a context, i.e. the name of the table, the names of category attributes, the corresponding subordinate categories and the names of the summary variables are given.
The set E of edges reflects the linking of nodes of type (X, X), (X, C), (C, C), and (C,X). There are some special features of this data model:

- the tree is representing more the physical structure of a statistical table (i.e. stab and heading of a table are modelled as subtrees).
- the tree is asymmetric in its subtrees. The right subtree models the heading of the table and refers columnwise to the data.

A following set of SUBJECT operations is available:

- **Browsing** is achieved by traversing the graph
- **Search** is provided to locate the file nodes directly using specific keywords
- **Examine** locating of nodes that contain specific keywords
- **Include** allowing for the specification of predicate conditions for queries
- **Aggregation** aggregating a selected set of terminal nodes
- **Display** displaying the result of a query in a table form
- **Document** displaying the text document associated with a node

3.2 NF^2-table structures

Ozsoyoglu and Yuan (1987) proposed a non-first-normal form (NF^2) of a nested relationship type. This form is called a multi-dimensional or multi-way table in statistics, cf Nelder (1974). It is a more natural representation of a complex data-structure because it is not »flattened out« like a normalized relationship. It is characterized by a crossing and nesting of the categorical attributes which together make up the stub and heading of the table. Each cell contains the corresponding value of the summary attribute linked to the relationship type. For example, consider the table »Professional Position in California« published in Ferri, Pisano, Rafanelli (1992) and printed in a slightly modified form below. It shows the absolute frequency distribution on the category attributes sex, year, professional category and qualification. The attributes sex and year are nested while sex and professional category are crossed.

3.3 *STORM* »Statistical Object Representation Model«

Rafanelli, Shoshani(1990) proposed a graph-oriented data model which is an enhanced version of the *SUBJECT* model. The history of this model is as follows:

SUBJECT (1981) → *GRASS* (1983) → *STORM.*

				Professional Category				
				Engineer		Secretary		Teacher
				Chemie. Engineer	Civil Engineer	Junior Secretary	Executive Secretary	Element. Teacher
			1981					
	Male		1982					
		Year						
			1988					
Sex			1989					
			1981					
	Female		1982					
		Year						
			1988					
			1989					

Tab.l: Number of professional positions in California (in 1000)
Source: Ferri, Pisano, Rafanelli (1992)

A complex data structure is used which describes a statistical object(StO). It is defined by the quadruple

StO = (N, Ca, S, f)

where
N is the name of the statistical object
Ca is a (finite) set of category attributes
S is a (single) summary attribute
f links Ca and S by a root tree.

A *STORM* model is a directed, acyclic graph(DAG) consisting of several topic nodes (T-nodes). A T-node is either a root node, or a node with at least one proceeding T-node and/or at least one succeeding S-node. A S-node is the root node of subtree corresponding to a single StO.
Allowing for

T-nodes	.... topic nodes
S-nodes	summary attribute nodes
X-nodes	cross product nodes
C-nodes	cluster nodes

a StO is represented by a *STORM* tree and a *STORM* model by a DAG.

A specific feature of *STORM* is to model non-balanced (»non-symmetric«) and heterogeneous statistical objects, cf. Rafanelli (1991). Non-symmetry arises when one category attribute is classified in a classification hierarchy (nomenclature), in which the number of levels is different depending on the category attribute referred to. For example, state, county and city form a hierarchy. Evidently, there exist states having cities but missing counties. Non-homogeneity arises when the instances of a category attribute are, in turn, classified with regard to different criteria. For example, the category attribute 'Professional Category ' has categories 'Engineer', 'Secretary' and 'Teacher'. While 'Teacher' may have subordinate categories like elementary, grammar or high school teacher, 'Engineer' may be subclassified according to the diploma degree etc.

3.4 **CSM** »Conceptual Statistical Model«

The data models presented so far were concerned with macro- and metadata. Di Battista and Batini (1988) introduced a model which covers all kind of data, however, using different paradigms.

1. Microdata or elementary data are represented by an ER-model
2. Macrodata or summary data are represented by a graph-based model similiar to the *SUBJECT* data-model.
3. Metadata are embedded in the ER-model and in the nodes of model graph representing the macrodata.

The conceptual schema of the macrodata is modelled by a labeled and marked DAG. Its nodes have the following semantic (Tab. 2). The corresponding pro-

node type	type of abstraction represented
S	Class of Statistical Object
C	Category Attribute
X	Statistical Classification
D	Class of Data
V	Data View
A	Aggregate
G	Grouping

Tab.2: The node types of the CSM data model

auction rules used for designing a feasible DAG are of the type »if a node is of type A then it has as parents nodes of type C and/or A«. For further details see Di Batista and Battini (1988).

3.5. **SDM4S** »Statistical Data Model based on a 4 Schema Concept«

This data model is devoted to macro- and metadata. The conceptual approach is strictly object-oriented. It has been developped and improved by Sato (1986, 1991) over a couple of years. Similiar ideas have been developped by Sundgren (1989, 1993) in order to improve the semantics of data modelling. He introduced the concepts of an object graph with formal definitions in an »infological« language called INFOL, of a metaobject graph with type, series and occurrence layers and of the so-called alfa-beta-gamma-tau-analysis, cf Sundgren (1991), which represents metadata from a substantial, regional and temporal point of view. The main features of Sato's model are the following:

- The model considers only macro- and metadata.
- The macrodata are stored physically in a relational database.
- The metadata are embedded in a statistical data dictionary (StDD). Its logical structure is described by an object-oriented (frame-oriented) system with 4 levels.

There are three classes of objects forming the data-dictionary frames.

- Statistical object
- Category value (domain)
- Summary value (domain).

The hierarchical structure is mapped by arcs of the type

- a_kind_of representing a superclass-subclass relationship,
- is_a representing a class-instance relationship,
- a_part_of representing a whole-part relationship.

Note, that the definition of the »is-a« relationship is rather unusual.
The main manipulating facilities are

- an editorfor insert, update and delete operations
- a browseruses the links between frames to explore the StDD and to list its items.

The statistical data-dictionary of the SDM4S is represented by a frame or object-oriented system. It consists of 3 types of classes (frames) and has 4 levels of abstraction.

(1) **The data model level** (root level of the frames):

There exist 3 root levels :
- Statistical objects categorized by
  - the categorical attributes with category values
  - and the summary attributes with summary values
- Category Values (Domain)
- Summary Values (Domain)

(2) **The Conceptual Level:**

On this level the data is described which is conceptually obtainable in the realm (the real object world) of a database regardless of its availability.

> **Ex.:**
> Persons are categorized by sex, age and summarized bythe population size, i.e by counting the number of persons classified according to their sex and age group.
> Persons, Employees represent statistical objects on the conceptual level
> Age and Sex Category,
> Population Size represent conceptual domains

(3) **DB Schema Level:**

On this level one describes which part of the conceptual data is available as actual and stored data in the StDB. There may be gaps due to sparse data. The distinction between concetually obtainable and actual available data is due to Sato(1991, p. 181). This idea makes it possible to give second best answers to a query. For example, a response could be »Census data on persons not available on a annual basis but on a 10 years basis!«.

Ex.:

Persons for Census are categorized by
sex = male/female
age = 5 years age group and summarized with
population = number with unit = 1000.
Persons for Censusrepresent an actual statistical object
5 years age grouprepresent an actual domain

(4) Instance Level

On this level the individual metadata and individual values are described which are linked to the objects on the instance level.

Ex.:

The metadata corresponding to a cell of a multi-dimensional table are kept here together with the name of the statistical object and the values of the categorical and summary attributes.

4. A M³ Database Architecture

As we pointed out in the preceeding chapters a statistical database must be represented conceptually by a three level architecture:

- microdata representing census, surveys or other collected data and stored at least as one rectangular scheme (»data matrix«) where rows represent cases(instances) and columns represent variables (attributes).
- macrodata representing aggregated or grouped data derived from microdata and stored in one or more multi-dimensional tables. At least one summary variable must be present like counts, average, index, quotient.
- metadata defining and describing micro- and macrodata on the
  - conceptual level
  - external level
  - internal level.

The conceptual level includes

- the semantic metadata (-> the meaning, the definitions, the substantial and legislative aspects of the data and its context)
- the statistical metadata (-> the sampling procedures, the estimation methods, transformation procedures already applied, datatypes in the statistical sense, processing technique of error bounds).

The external level includes the database view-oriented metadata of, cf. van den Berg et al. (1992):

- **the data suppliers** (-> on the census, survey, panel or reporting schema and frames, the definitions of classes and instances, the corresponding attributes and relationsships like nomenclatures as well as specific characteristics of the generation process including footnotes and documents)
- **the data producers** (-> on methodology and rules used for inputting, chekking, transforming, producing and storing and retrieving data and metadata)
- **the data users or consumers** (-> on retrieving, outputting and interpreting data including the granting of the access rights and the link to the logbook).

The internal level includes the storage metadata like file access techniques and file data protection and data security aspects.

This kind of metadata should not be mixed up with the metadata stored in the data dictionary (DD) of a database management system. The metadata in the DD represents the more technical metadata like names of identifiers, names of attributes, aliases, datatypes, lengths, key/non key-characteristics of primary and secondary type of keys, null/not null-entries etc Evidently, it is necessary to have a non technical supplement of the metadata on the conceptual, internal and external levels. A suitable datamodel for this kind of data is a thesaurus (glossar with a cross-referencing) linked to a document retrieval system.

4.1 Microdata

Microdata are the primary or atomic data on individuals, objects, events etc. They come from various sources and are collected by a census, survey, panel, report etc. In a standardized form they are linked to one and only one class of instances. For example, the German 1987 census was aiming for persons, housing and jobs. Consequently, the data are splitted into the three classes 'Person', 'Housing' and 'Job'.

As is specifically true for sampled, reported or census data, the datatype »data-matrix« is almost suitable because of its »flat form«. The rows represent the cases and the columns the variables. Note, that the case identifier is almost not concatened, i.e. the microdata is indexed in simple way. For example, the case number is a scalar case identifier (primary key). This feature is opposite to the macrodata, where the key is multi-dimensional or concatenated, i.e. the data set is highly indexed.

The term »data-matrix« is misleading although it is rather popular among statisticians. The columns of a »data-matrix« are heterogenous with respect to the datatype like boolean, cardinal, integer, decimal, real, string etc. Therefore, the term »relation« or »(flat) table« is more appropriate for such a datatype. It can formally be defined as follows:

(1) **data structure (scheme):** a relationship (flat table) scheme R of degree (k+1) between k data attributes

(R_Name) ((primary_key), (data attribute$_1$)
.
.
.
(data attribute$_k$))

(2) **functional dependency:** if the two tupels (rows) t_i, $i \in R$ have the same value of the primary key then the values of the data attributes 1 ... k are identical.

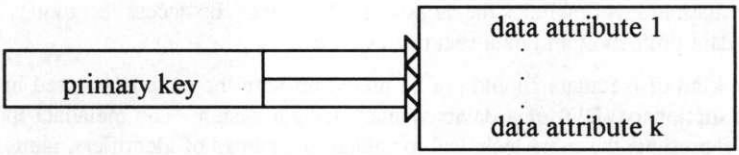

Fig.2: The functional dependency structure of microdata

(3) **Operations on a relation R:**

Π_s^R projection of R on a subset s of columns (attributes)
σ_c^R selection of rows (tuples) from R according to the selection criteria c
$R_1 \rhd\lhd R_2$ join of the two relations R_1, R_2
$\cap, \cup, \setminus$ set operations on R_1, R_2
$O(x) = x\ O(x-1)$ recursive query on R.

The data model »datamatrix« has the following specific features:
- Unique up to permutation of rows and columns
- Homogeneous datatypes columnwise
- Weak model for time-series data
- Scalar primary key (case identifier).

The metadata linked to microdata includes:
- the name of the relationship type (the statistical object class) and its definition
- names,datatypes and domains (and/or codes) of the attributes together with its definition, nomenclature and the key/non key and null/not null values characteristics
- the coded/not coded value labels of the category attribute
- the periodicity of a temporal, category attributes

- scale, dimension, unit, stock/flow type, status, error rate etc. of a summary attribute
- the source of data with documents on the institutional, legislative and statistical background
- the privacy rules.

4.2 Macrodata

Generally speaking, macrodata are data derived from microdata applying operations like grouping and aggregation with respect to subject, time and space. In a standardized form provision is made to link the macrodata to one and only one statistical class or object type like people, job, housing, industry, elections etc. This guarantees that specific macrodata are referring to one statistical population only and improves the flexibility of the database design. The data are highly indexed, i.e. there is at least one summary attribute which is fully dependent on the whole set of the corresponding category attributes.

(1) **data structure (scheme):** a multi-dimensional table T formed by p categorial attributes and at least one summary attribute referring to a single statistical set or population (universe).

(TName) (categ. attr.$_1$, ... , categ. attr.$_p$, summary, attr.)

(2) **(foil) functional dependency:** the summary attribute is functional dependent on the whole set of category attributes and not on a proper subset of it.

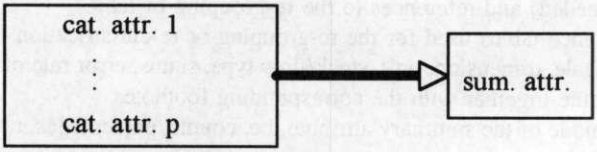

Fig.3: The functional dependency structure of the macrodata

Note that a p-dimensional table T can be represented by a relation R' of degree (p+1) using the p category attributes as a concatenated primary key and the summary attribute as the only non-key attribute, i.e.

R': <RName> (<primary key>, summary attribute)

However, the relational datamodel flattens out the category attributes and seems therefore to be neither conceptually convincing nor computational efficient.

(3) **Operations**

The data structure of a »multi-dimensional table« is much more complex one than the the simple »datamatrix« of the microdata which can be represented by a relational datamodel. The whole set of relational operations is inappropriate and must be re-defined for macrodata, cf. Meo-Evoli, Ricci, Shoshani (1992).

- table arithmetic (+, -, x, / - join)
- table-union
- table-selection (»conditioning«)
- table-projection (»marginalization« or »summarization«)
- table-reclassification
- table extension.

(4) Metadata

The metadata needed to describe and define conceptually the macrodata on the statistical and semantic level are the following ones:

- the name of the statistical object and its corresponding population together with its definition
- the questionnaire or report form used for surveys, census, panels or regular reports together with the statistical documentation and the legislative foundation
- the names, datatypes, domains, co-domains and definitions of the summary attribute and the category attributes together with the nomenclatures (hierarchies within the category attributes)
- the value labels (coded/not coded) of the category attributes with footnotes (if needed) and references to the re-grouping of tables
- reference tables used for the re-grouping or re-classification
- the scale, dimension, unit, stock/flow type, status, error rate of the summary attribute together with the corresponding footnotes
- the mode of the summary attribute, i.e. counts, percentages, ratios, averages etc.)

4.3 Modelling metadata using an eER-diagram

A comprehensive formal way to model metadata is to use an extended entity-relationship model (eERD). An eER diagram is a graph which consists of

the nodes of type »rectangulars« representing entity types or classes of meta-objects which have similiar characteristics

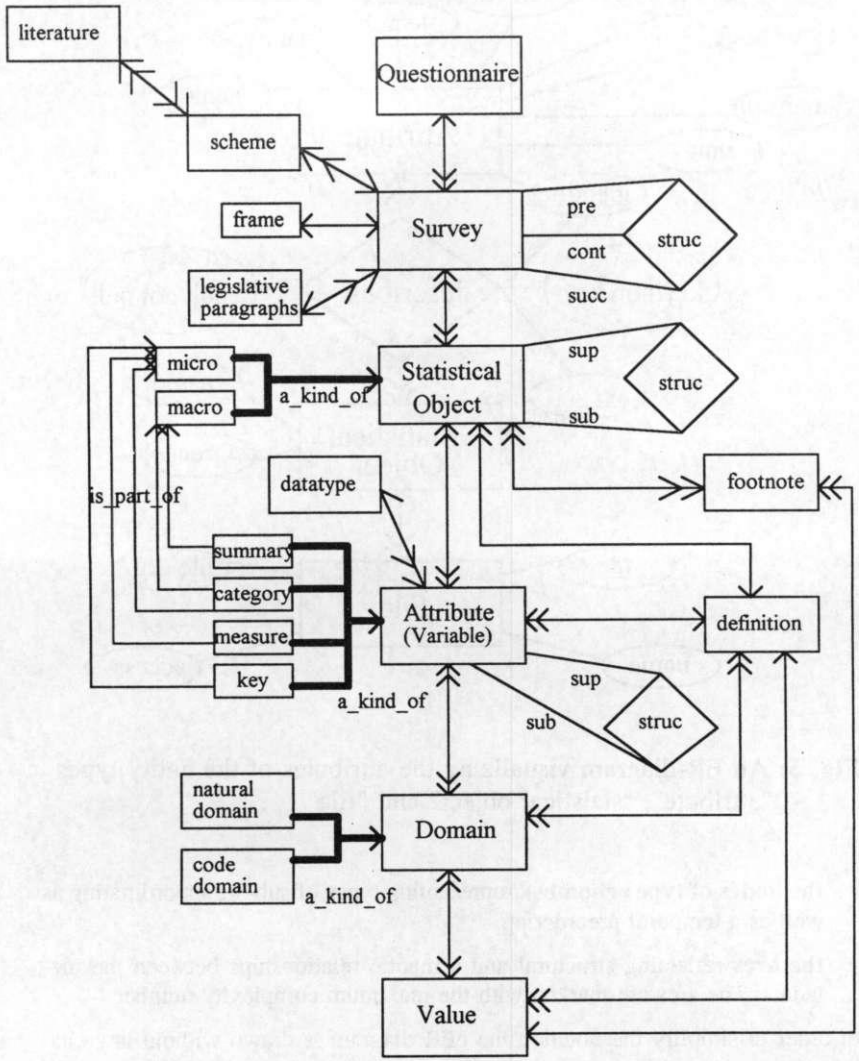

Fig.4: An extended ER-diagram for metadata

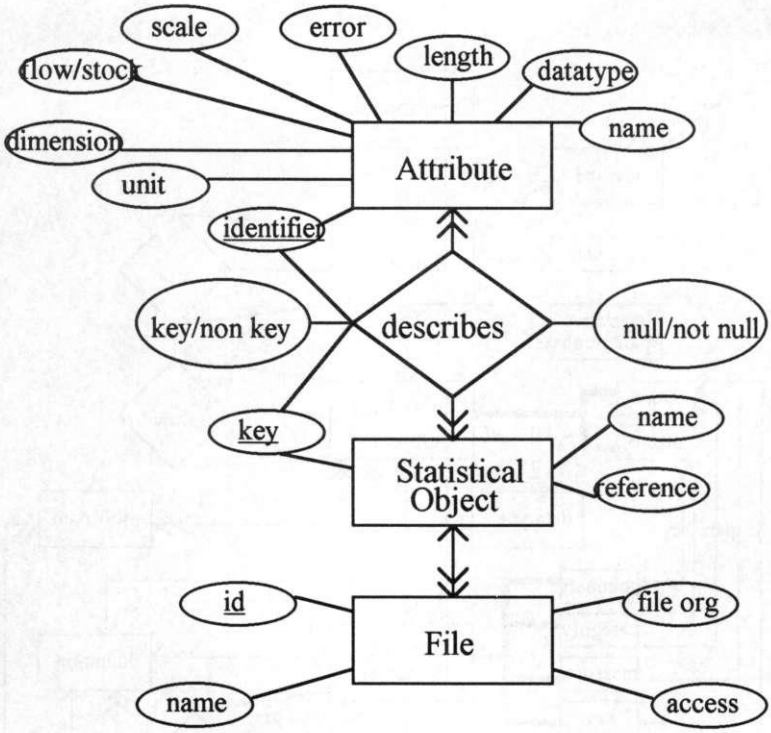

Fig. 5: An ER-diagram visualizing the attributes of the entity types "attribute", "staistical object" and "file"

the nodes of type »rhombs« representing types of sub- or supordinating as well as a temporal preordering

the arcs reflecting structural and semantic relationships between the metadata. The arcs are marked with the maximum complexity number.

In order to simplify the notation the eER diagram is drawn without any characteristics, i.e. attributes of the entity types of the metadata. Of course, a detailed structure can be visualized by a stepwise refinement of each subset of entity types or classes. This is demonstrated below (Fig. 4).

The ER-diagram shown below (Fig. 5) is visualizing die attributes of the (meta) entity types »attribute«, »statistical object« and »file«. It should be considered as an example for a further refinement of an eER-diagram. it reflects the semantic, statistical and storage views on the metadata.

The operations needed to create, update and explore the metadata are

- **loading, editing a metadata graph**
- **scrolling**
- **browsing**
- **searching**
- **zooming**
- **listing, printing, storing.**

References

Di Battista, G. and Batini, C.(1988), Design of Statistical Databases: A Methodology for the Conceptual Step, Inform. Systems, Vol.13, No. 4, pp. 407-422.

van den Berg, G.M., de Feber, E., and de Greef, P. (1992), Analysing Statistical Data Processing, in Eurostat, New Technologies and Techniques for Statistics, Luxembourg, pp 102-111.

Chan, P., and Shoshani, A. (1981), SUBJECT: A Directory Driven System for Large Statistical Databases, in Proc. of the LBL Workshop on Statistical Database Management, Lawrence Berkeley Lab, Berkeley, CA.

Cubitt, R. et al.(eds.) (1986), Proceedings of the Third International Workshop on Statistical and Scientific Database Management, Eurostat, Luxembourg.

Eurostat (1992), New Technologies and Techniques for Statistics, Luxembourg.

Ferri, F., Pisano, M.T., and Rafanelli, M.(1992), A Object Oriented Visual Definition Language for Statistical Data, in New Technologies and Techniques for Statistics, Proc. of the conference, Bonn, pp.320-339.

Hinterberger, H., French, J.C.(eds.) (1992), Proceedings of the Sixth International Workshop on Statistical and Scientific Database Management, Department of Informatik, ETH Zürich, Zürich.

Lenz, H.-J.(1993), M3-Database Design, Micro-, Macro- and Metadata Modelling, in F. Faulbaum (ed.), SoftStat '93 Advances in Statistical Software 4, Gustav Fischer, Stuttgart etc, forthcoming.

Meo-Evoli, L., Ricci, F.L., and Shoshani, A. (1992), On the semantic completeness of macro-data operators for statistical aggregation, in Hinterberger, H., French, J.C.(eds.), Proceedings of the Sixth International Workshop on Statistical and Scientific Database Management, Department of Informatik, ETH Zürich, Zürich, pp. 239-258.

Michalewicz, Z. (ed.) (1990), Proceedings of the Fifth International Workshop on Statistical and Scientific Database Management, Charlotte, NC, Lecture Notes in Computer Science, vol. 420, Springer Verlag, New York etc.

Michalewicz, Z. (ed.) (1991), Statistical and Scientific Databases, Ellis Horwood, New York etc.

Neider, J.A.(1974), Genstat - A Statistical System, COMPSTAT 1974, Proc.in computational statistics, Bruckmann, G. et al (eds.),Physica Verlag, Wien, pp.499-506.

Ozsoyoglu, G and Yuan, L.Y. (1987), A New Normal Form for Nested Relations, ACM Transactions on Database Systems, Vol.12, No. 1.

Rafanelli, M. et al. (eds.) (1988), Proceedings of the Fourth International Workshop on Statistical and Scientific Database Management, Roma, Lecture Notes in Computer Science, vol.339, Springer Verlag, New York etc.

Rafanelli, M., and Shoshani, A. (1990), STORM: A Statistical Object Representation Model, in Michalewicz, Z. (ed.), Proceedings of the Fifth International Workshop on Statistical and Scientific Database Management, Charlotte, NC, Lecture Notes in Computer Science, vol. 420, Springer Verlag, New York etc, pp. 14-29.

Sato, H. (1986), Conceptual Schema for a Wide-Scope Statistical Database and its Applications, in Cubitt, R. et al.(eds.), Proceedings of the Third International Workshop on Statistical and Scientific Database Management, Eurostat, Luxembourg.

Sato, H. (1991), Statistical Data Models: from a Statistical Table to a Conceptual Approach, in Michalewicz, Z. (ed.), Statistical and Scientific Databases, Ellis Horwood, New York etc., pp. 167-200.

Shoshani, A. (1982), Statistical Databases: Characteristics, Problems, and some Solutions, Proc. of the 8th Int. Conf. on very Large Data Bases (VLDB), pp.208-222.

Sundgren, B. (1989), Conceptual Modelling as an Instrument for Formal Specification of Statistical Information Systems, ISI 47th Session, Paris.

Sundgren, B. (1993), Modelling Meta-Information Systems, in Statistical Meta Information Systems, Proc. of the Conference, Office for Official Publications of the European Communities, Brussels, Luxembourg, pp. 59-79.

Ullman, J.D. (1988), Principles of Database and Knowledge-Base Systems, Vol. I, Computer Science Press, Rockville World Systems (1992), Conference Proceedings, Statistical Meta Information System Workshop, Luxembourg.